BRITISH & FOREIGN BIBLE SOCIETY

Hunting the Snark

The problem posed for MT by non-concatentive morphologies

They sought it with thimbles,
They sought it with care,
They pursued it with forks and hope,
The threatened its life with a railway share,
They cajoled it with smiles and soap.

Linguistic Computing at BFBS
linguistics@biblesociety.org.uk
http://lc.bfbs.org.uk

BRITISH & FOREIGN BIBLE SOCIETY

A common question

- Why not use Google Translate?
  - 65 languages
  - saves time & money
  - etc...

BRITISH & FOREIGN BIBLE SOCIETY

Genesis 1.1

- Ancient Greek - LXX
  - ἐν ἀρχῇ ἐποίησεν ὁ θεὸς τὸν οὐρανὸν καὶ τὴν γῆν. ἡ δὲ γῆ ἦν ἀόρατος καὶ ἀκατασκεύαστος καὶ σκότος ἐπάνω τῆς ἀβύσσου καὶ πνεῦμα θεοῦ ἐπεφέρετο ἐπάνω τοῦ ὕδατος.
- English...
  - in the beginning epoíisen God heavens and the earth. And the earth unseen and HN akataskeúastos and Scott epáno avússou and Spirit of God epeféreto epáno waters.

BRITISH & FOREIGN BIBLE SOCIETY

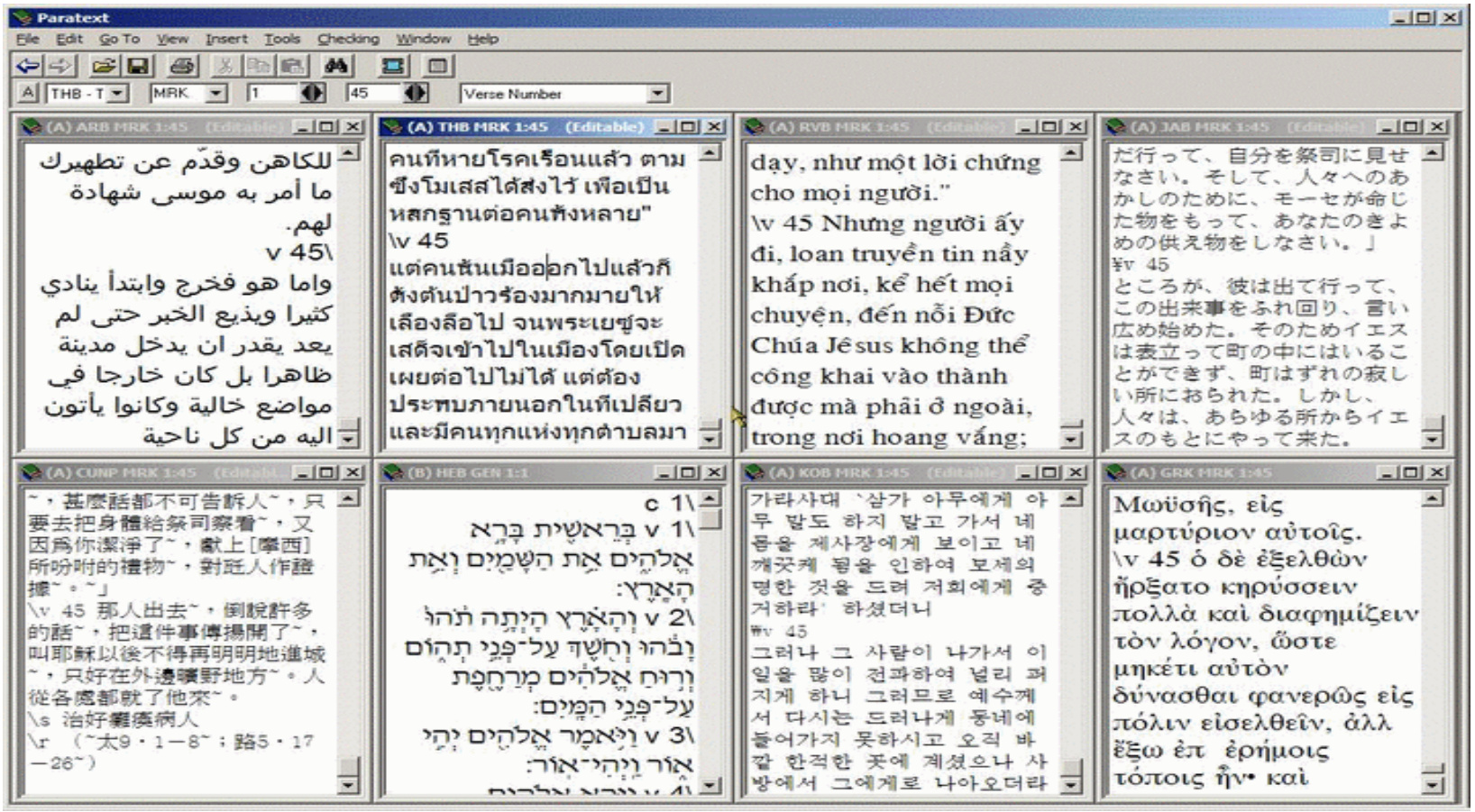
One or two problems

- Fixable
  - better Gk lexicon
  - and grammar
  - i.e. put in more knowledge
- Harder to fix
  - Target language
    - any of 7,000 +
    - no lexicon
    - no grammar
    - different language tomorrow...

BRITISH & FOREIGN BIBLE SOCIETY

ParaTExt

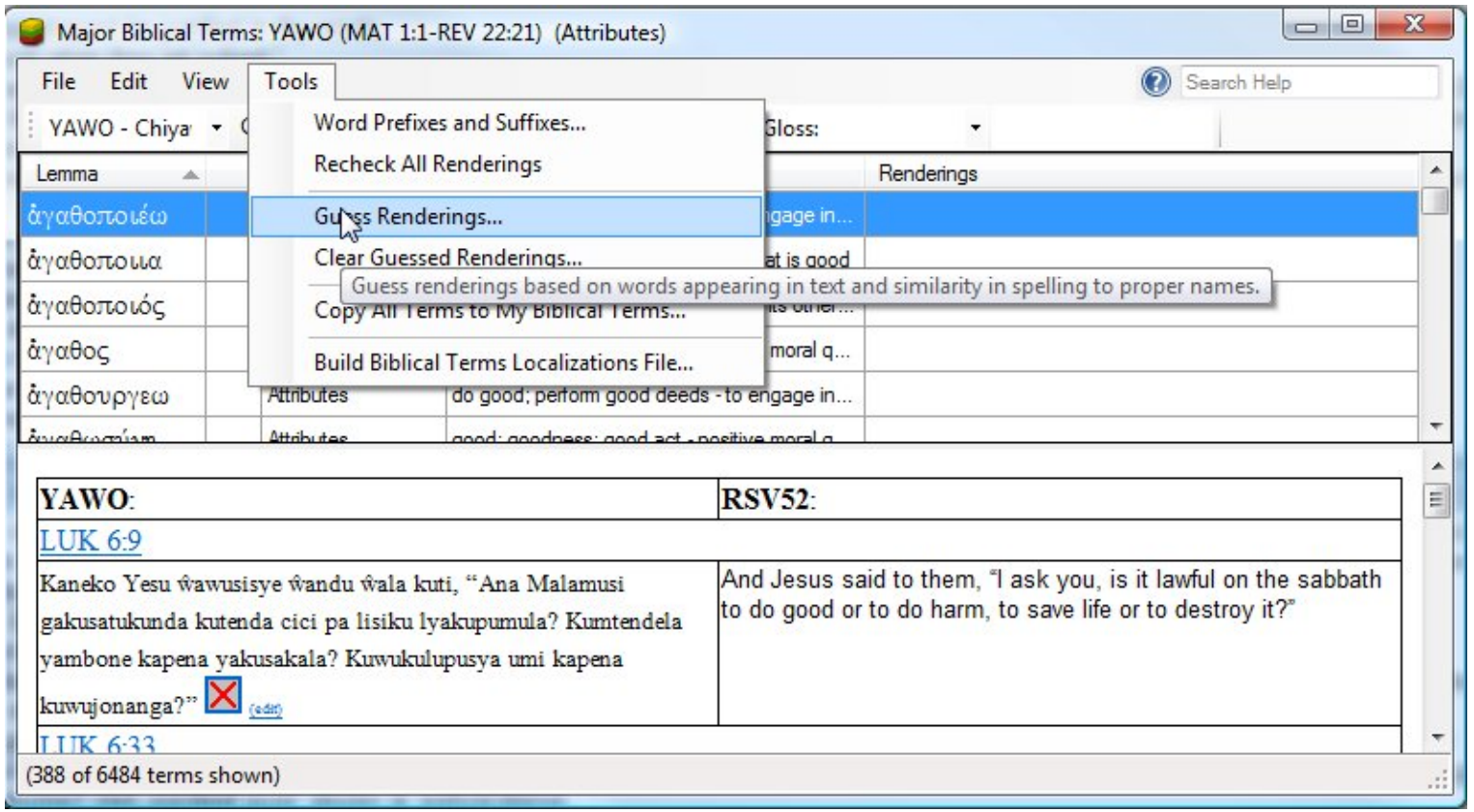
- A translator's workbench (language independent)
  - Editing environment, Base texts, Model texts



BRITISH & FOREIGN BIBLE SOCIETY

Checking & Review

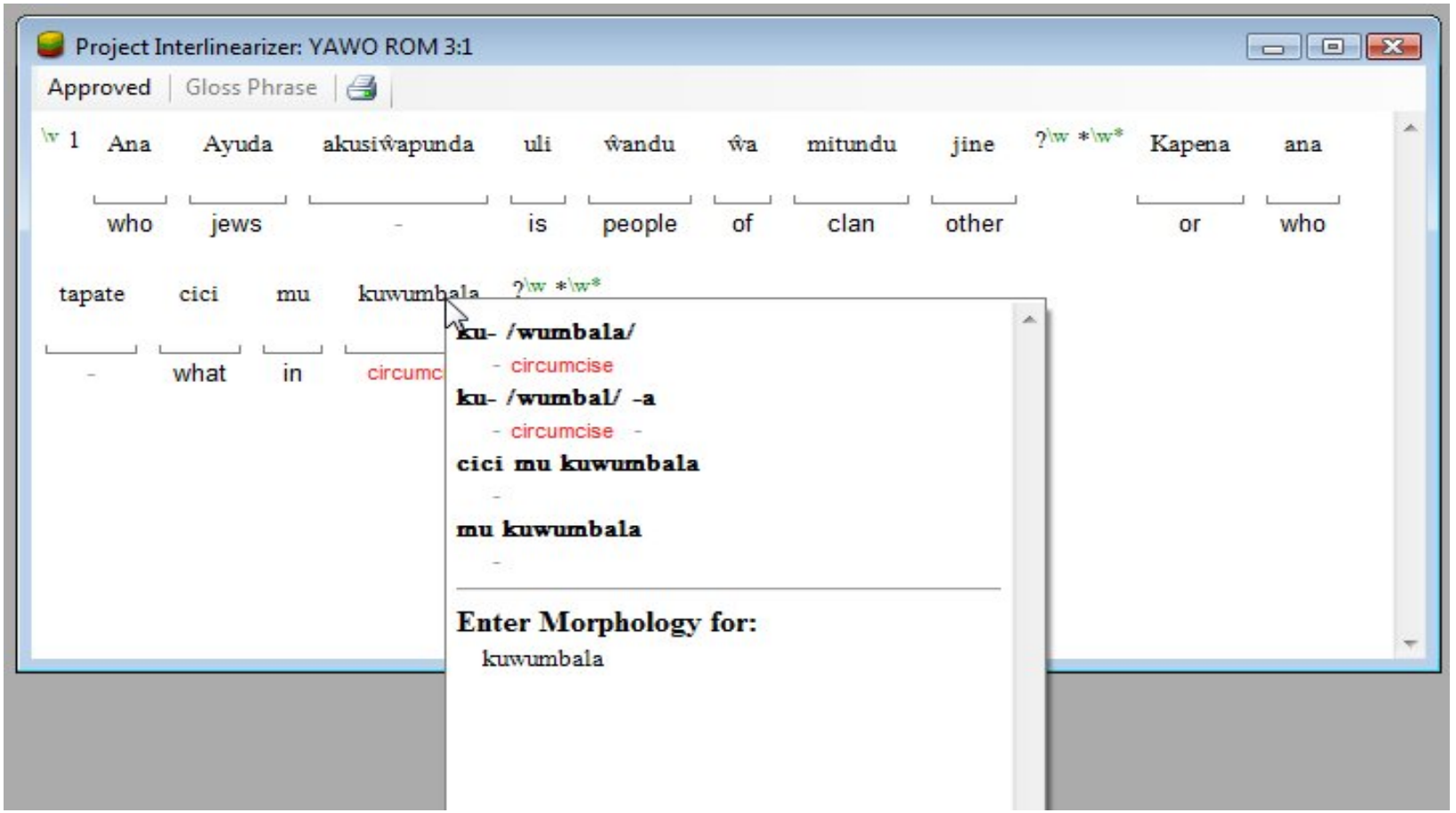
- Key term list
  - Automatic analysis of text for consistency
- Semi-automatic morphology analysis
  - Spelling checks



BRITISH & FOREIGN BIBLE SOCIETY

Checking & Review

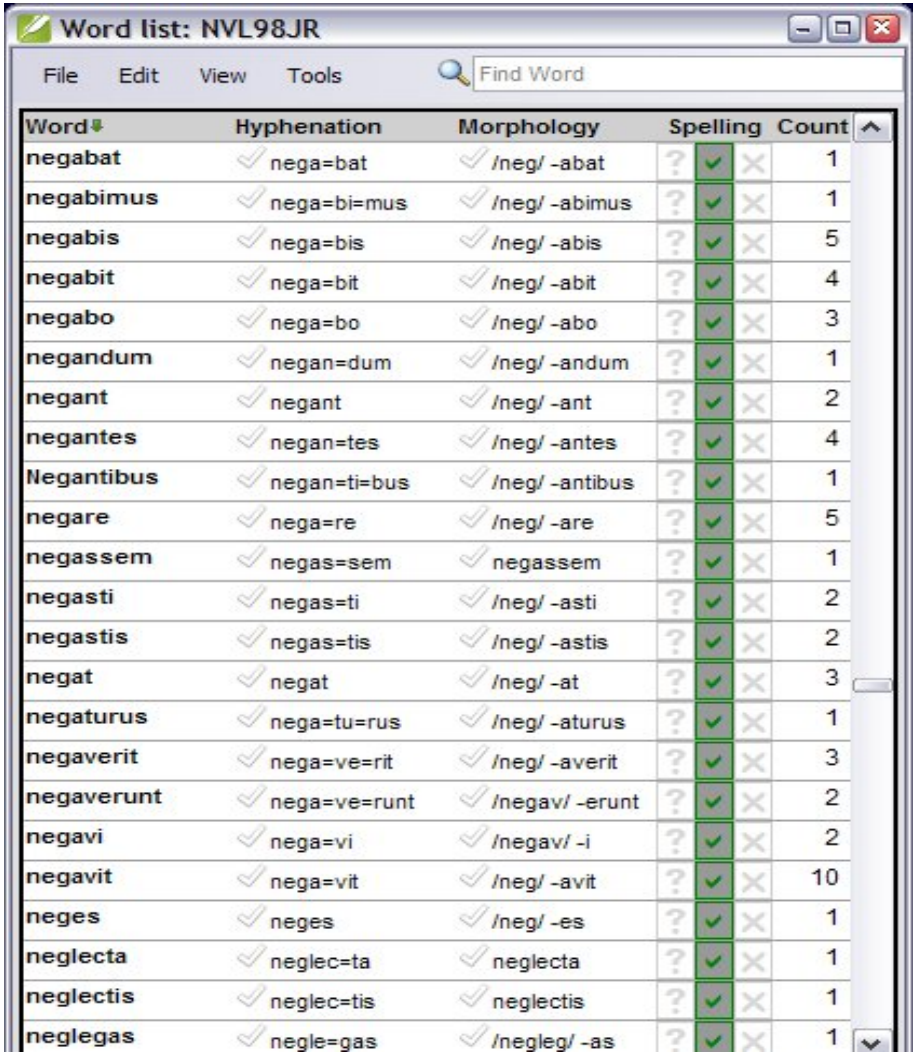
- Automatic interlinear back-translation



BRITISH & FOREIGN BIBLE SOCIETY

Spelling review

- Word list tool
  - parsed by
    - morphology
    - syllable

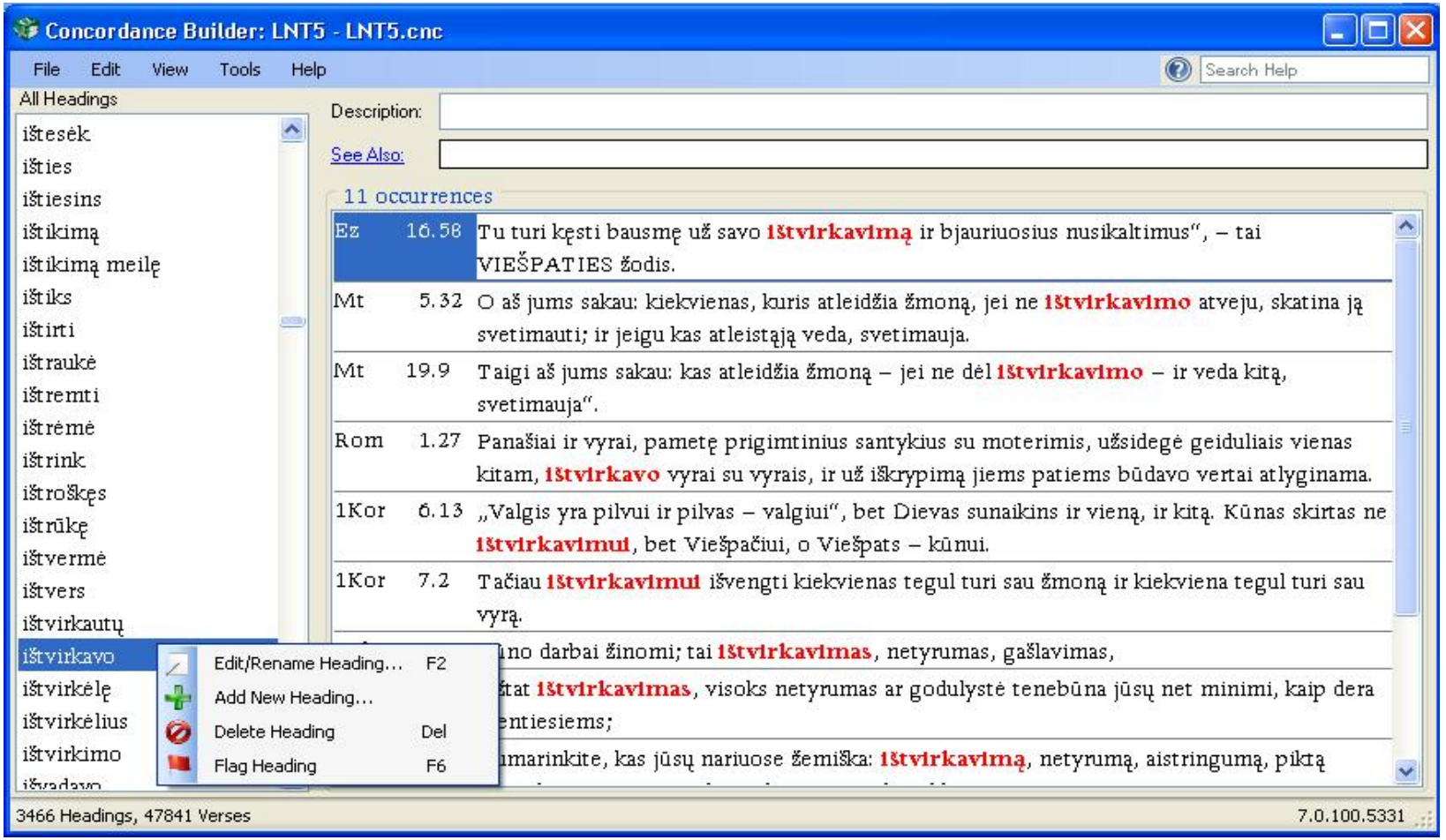


Word	Hyphenation	Morphology	Spelling	Count
negabat	nega=bat	/neg/ -abat		1
negabimus	nega=bi=mus	/neg/ -abimus		1
negabis	nega=bis	/neg/ -abis		5
negabit	nega=bit	/neg/ -abit		4
negabo	nega=bo	/neg/ -abo		3
negandum	negan=dum	/neg/ -andum		1
negant	negant	/neg/ -ant		2
negantes	negan=tes	/neg/ -antes		4
Negantibus	negan=ti=bus	/neg/ -antibus		1
negare	nega=re	/neg/ -are		5
negassem	negas=sem	negassem		1
negasti	negas=ti	/neg/ -asti		2
negastis	negas=tis	/neg/ -astis		2
negat	negat	/neg/ -at		3
negaturus	nega=tu=rus	/neg/ -aturus		1
negaverit	nega=ve=rit	/neg/ -averit		3
negaverunt	nega=ve=runt	/negav/ -erunt		2
negavi	nega=vi	/negav/ -i		2
negavit	nega=vit	/neg/ -avit		10
neges	neges	/neg/ -es		1
neglecta	neglec=ta	neglecta		1
neglectis	neglec=tis	neglectis		1
neglegas	negle=gas	/negleg/ -as		1

BRITISH & FOREIGN BIBLE SOCIETY

Beyond translation

- Literacy - Dictionaries & Concordances



BRITISH & FOREIGN BIBLE SOCIETY

Glossing Technologies

- Provide
  - Language independent
  - Lemmatisation
  - Morphology analysis
    - driven by glossing
- Problems
  - Orthography
    - spaces...
  - Complex morphologies

BRITISH & FOREIGN BIBLE SOCIETY

Complex Morphologies 1

- Concatenative
  - >75% of languages
  - e.g. Bantu languages
    - Swahili
    - verb -pend-
  - Word Form Template:
    - **[Pre]Stem[Suff]**

- *akipenda, anakupenda, atanipenda, mlipenda, mpende, nakupenda, nawapenda, nilipenda, ninakupenda, [-]pendana, [-]pendea, [-]pendwa, sikupendi, tulipenda, tutapenda, ulipenda, ungependa, utapenda, walipenda, wanaupenda, watapenda*

BRITISH & FOREIGN BIBLE SOCIETY

Complex Morphologies 2

- Non-concatenative
  - <25%
  - e.g. Semitic languages
    - Amharic, Arabic, Hebrew, Syriac
  - Template:
    - [m]$[m]$[m]$[m]
      - $ = stem
      - m = morphs

קָטַל קָטַלְנוּ תִּקְטֹל
יִקְטְלוּ קִטְלִי
קְטוּלוֹת יִקְטְלוּ

QF+AL QF+AL:NW.
TIQ:+OL YIQ:+:LW.
QO+:L”Y Q:+W.LOWT
YIQ:+:L”HW.

BRITISH & FOREIGN BIBLE SOCIETY

Finding morpheme structures

- Premise:
  - Valid morpheme structures will occur in a text with statistically significant frequency
- Three ways to find morphemes
  - Statistically
  - Minimum Description Length
  - Paradigm analysis
    - i.e. by stem association

BRITISH & FOREIGN BIBLE SOCIETY

Paradigm Analysis

- Concatenative
  - Find possible morphs
    - examine initial and final n-grams
  - Validate
    - build inflection paradigms
- Non-concatenative
  - Find possible morphs
    - ?
  - Validate
    - build inflection paradigms

BRITISH & FOREIGN BIBLE SOCIETY

Finding morph templates

- Pre-requisites
  - a lexicon of surface forms in the target language
- Method
  - compare each form in the lexicon with every other form and note common sequences.

BRITISH & FOREIGN BIBLE SOCIETY

Hebrew - Rendering

- UTF-8
  - Difficult to render zero width glyphs
- Michigan-Claremont Enconding
  - 7-bit ASCII
  - Easier to render morph templates
    - Remove cantillation

Example MC Encoding:

UTF-8	MC
With Cantillation & Vowels:	
בְּרֵאשִׁית	B.:/R")$I73YT
Without Cantillation:	
בְּרֵאשִׁית	B.:/R")$IYT
Consonants alone:	
בראשית	BR)$YT

fig. 5.

BRITISH & FOREIGN BIBLE SOCIETY

Example 1 - QF+AL / MFLA+

- Four matched characters: MF+AL
- Rule:
  - successors must follow below and to the right
- Three solutions:
  - F+, FA, FL

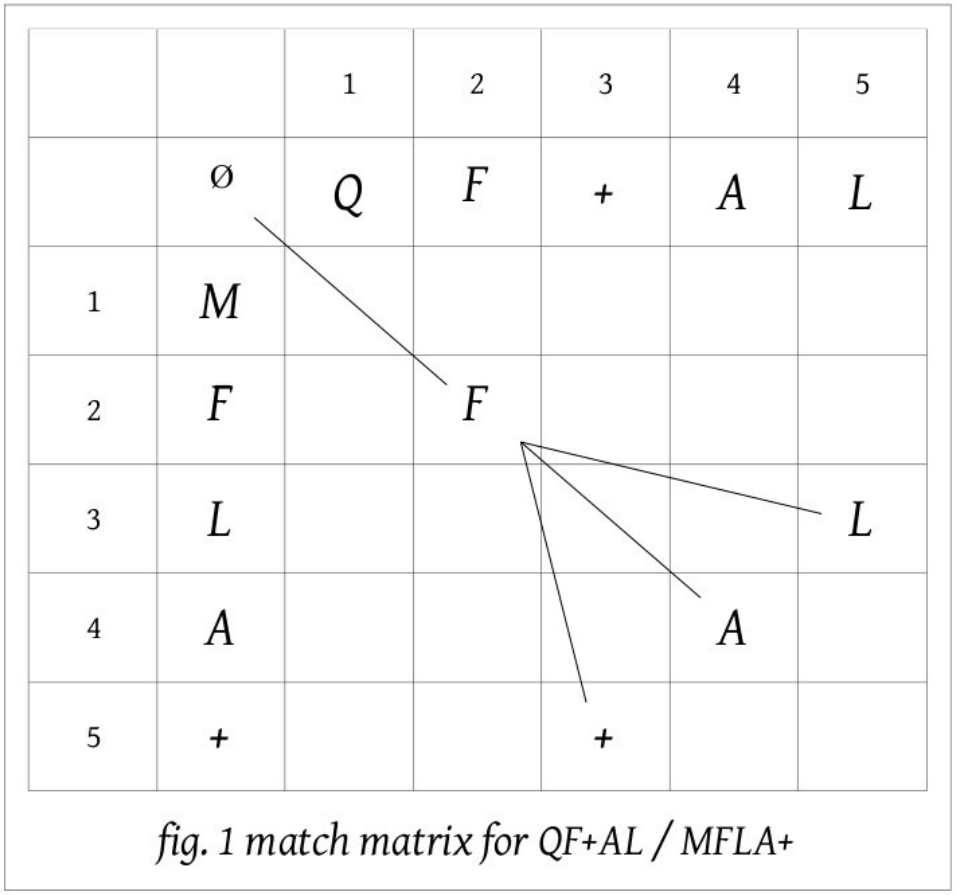


fig. 1 match matrix for QF+AL / MFLA+

Assessing the solutions

- Match coordinates:
  - F(2,2)
  - +(3,5)
  - A(4,4)
  - L(5,3)
- S1
  - { F(2,2) , +(3,5) }
- S2
  - { F(2,2) , A(4,4) }
- S3
  - { F(2,2) , L(5,3) }

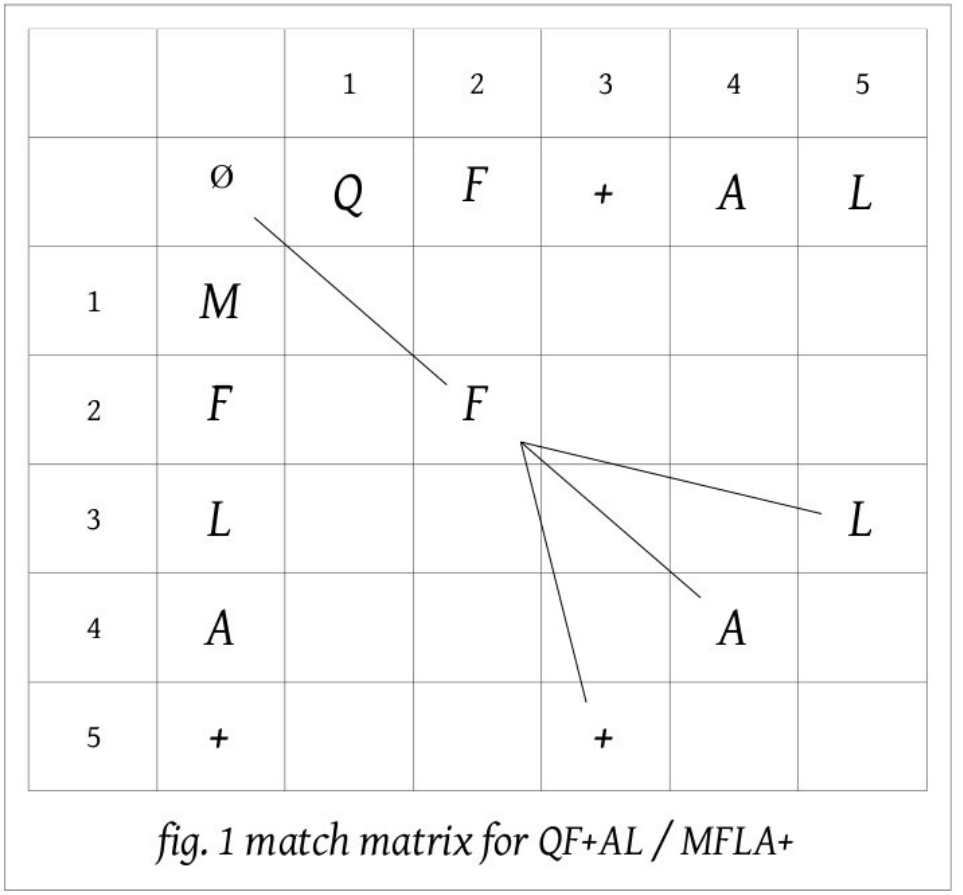


fig. 1 match matrix for QF+AL / MFLA+

BRITISH & FOREIGN BIBLE SOCIETY

Solution Value (V)

- V = 1 + (1 – d/f)
  - where
    - d = distance between the two x or y coordinates, whichever is the greater.
    - f = 10 (distance beyond which it is unlikely the two items are related)

- S1: F(2,2) , +(3,5)
  - V = 1.7
- S2: F(2,2) , A(4,4)
  - V = 1.8
- S3: F(2,2) , L(5,3)
  - V = 1.7

Example 2 - YIM:LO+ / YIQ:+OL

- Matched items:
  - Y (1,1)
  - I (2,2)
  - : (4,4)
  - L (5,7)
  - O (6,6)
  - + (7,5)

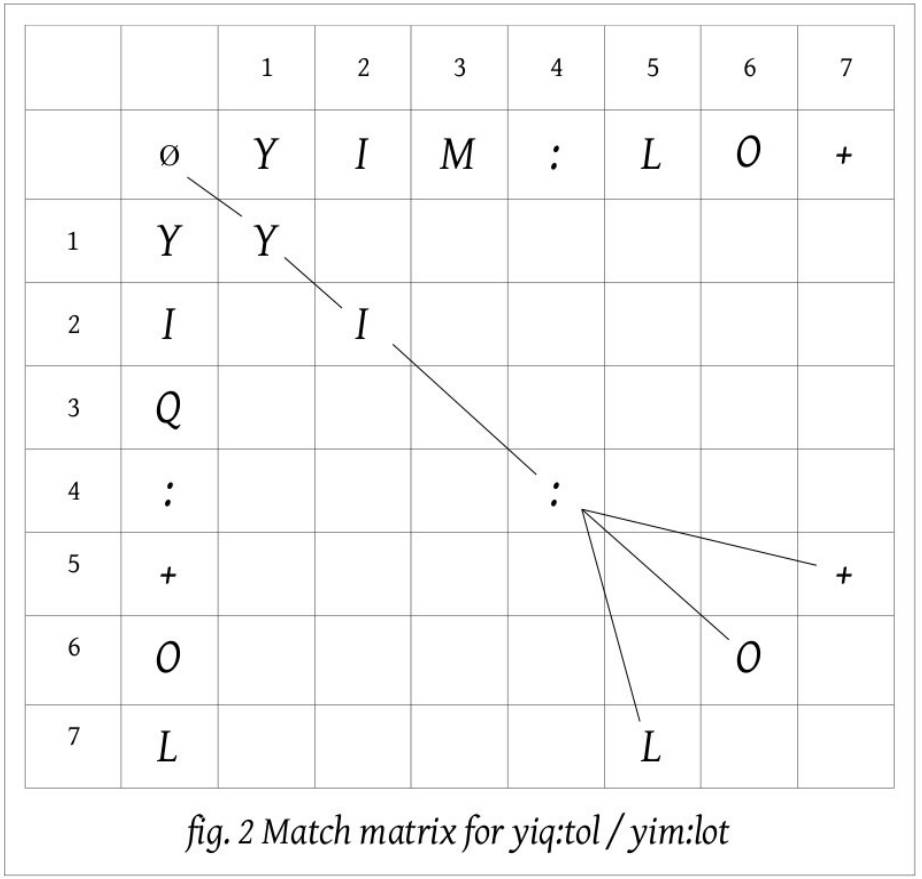


fig. 2 Match matrix for yiq:tol / yim:lot

Example 2 - Solutions

- S1 { Y(1,1), I(2,2), :(4,4), +(7,5) }
  - 1.9, 1.8, 1.7 = 5.814
- S2 { Y(1,1), I(2,2), :(4,4),O(6,6) }
  - 1.9, 1.8, 1.8 = 6.156
- S3 { Y(1,1), I(2,2), :(4,4), L(5,7) }
  - 1.9, 1.8, 1.7 = 5.814

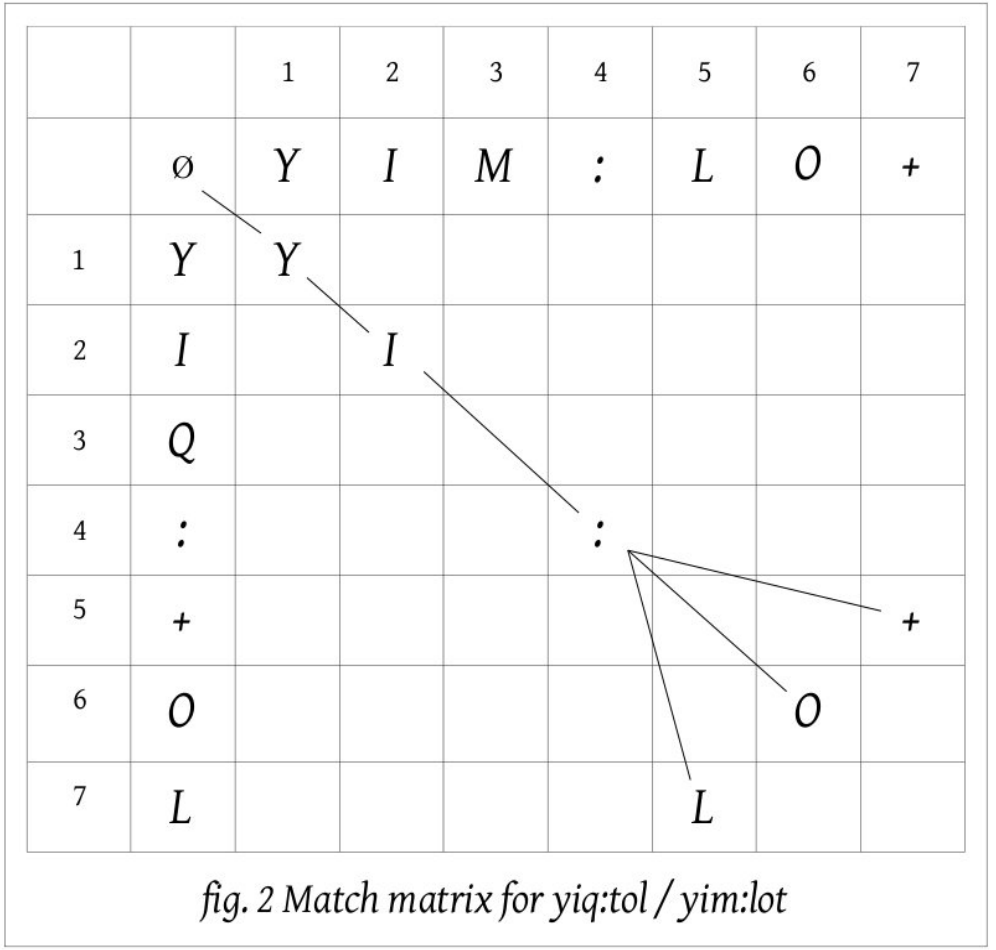


fig. 2 Match matrix for yiq:tol / yim:lot

BRITISH & FOREIGN BIBLE SOCIETY

Example 3 - YIQ:+:LW. / YIM:L:+W.

- Y(1,1)
- I (2.2)
- : (4,4)
- : (4,6)
- L(5,7)
- : (6,4)
- : (6,6)
- +(7,5)
- W.(8,8)

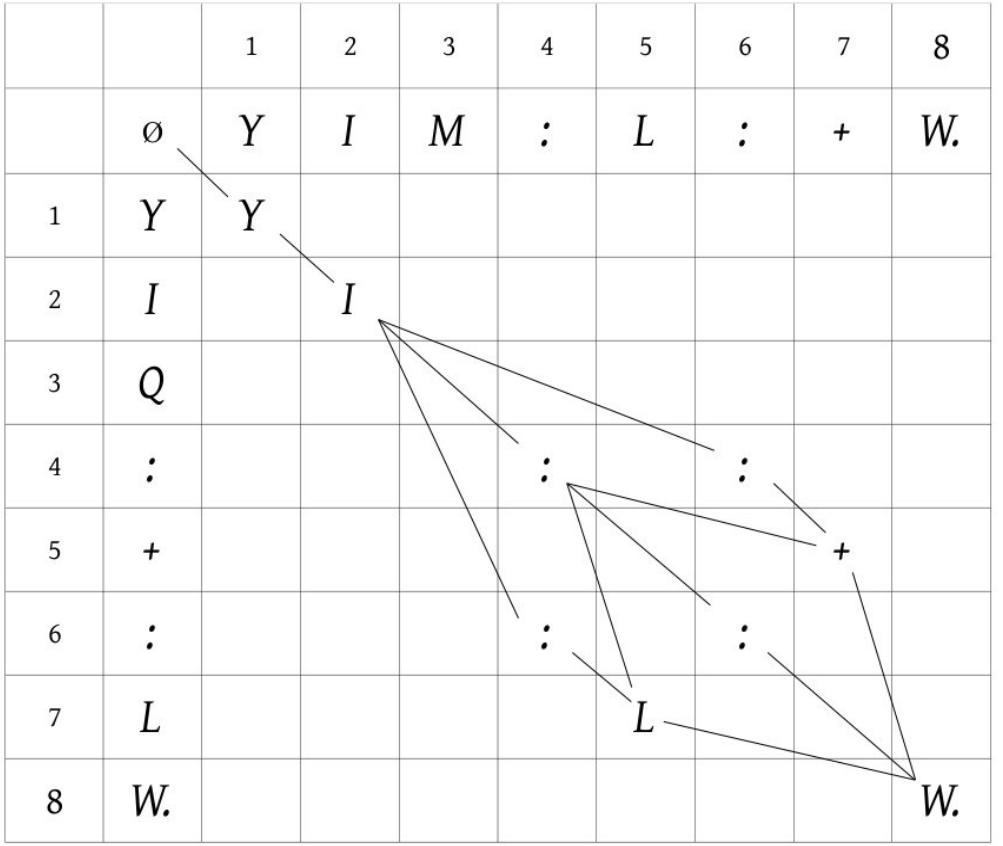


fig 3. Match matrix for YIQ:+LW / YIM:L:+W

BRITISH & FOREIGN BIBLE SOCIETY

E.g. 3 Solutions

- S1{Y(1,1), I(2,2), :(6,4), L(7,5), w.(8,8)}
  - 1.9•1.6•1.9•1.7 = 09.8192
- S2{Y(1,1), I(2,2), :(4,4), L(7,5), w.(8,8)}
  - 1.9•1.8•1.7•1.7 = 09.8838
- S3{Y(1,1), I(2,2), :(4,4), :(6,6), w.(8,8)}
  - 1.9•1.8•1.8•1.8 = 11.0808
- S4{Y(1,1), I(2,2), :(4,4), +(5,7), w.(8,8)}
  - 1.9•1.8•1.7•1.7 = 09.8838
- S5{Y(1,1), I(2,2), :(4,6), +(5,7), w.(8,8)}
  - 1.9•1.6•1.9•1.7 = 09.8192

BRITISH & FOREIGN BIBLE SOCIETY

Solution complements

- Eg. 1 QF+AL / MFLA+
  - S1{F(2,2), +(3,5)} Q__AL, M_LA_
  - S2{F(2,2), A(4,4)} Q_+_L, M_L_+
  - S3{F(2,2), L(5,3)} Q_+A_, M_L__
- Eg. 2 YIQ:+OL / YIM:LO+
  - S1 { Y(1,1), I(2,2), :(4,4), +(7,5) } __Q__OL, __M_LO_
  - S2 { Y(1,1), I(2,2), :(4,4), O(6,6) } __Q_+_L, __M_L_+
  - S3 { Y(1,1), I(2,2), :(4,4), L(5,7) } __Q_+O_, __M__O+
- Eg. 3 YIQ:+LW. / YIM:L:+W.
  - S1{Y(1,1), I(2,2), :(6,4), L(7,5), W.(8,8)} __Q:+___, __M__:+_
  - S2{Y(1,1), I(2,2), :(4,4), L(7,5), W.(8,8)} __Q_+:__, __M__:+_
  - S3{Y(1,1), I(2,2), :(4,4), :(6,6), W.(8,8)} __Q_+_L_, __M_L_+_
  - S4{Y(1,1), I(2,2), :(4,4), +(5,7), W.(8,8)} __Q__:L_, __M_L:+_
  - S5{Y(1,1), I(2,2), :(4,6), +(5,7), W.(8,8)} __Q__:L_, __M:L___

BRITISH & FOREIGN BIBLE SOCIETY

Hebrew Results

- Lexicon
  - Hebrew forms in Genesis – 4,431
- Templates Generated
  - 52, 357
- Best 1% of templates
  - Complements
    - 50 stems,
    - 42 valid
- Build inflection paradigms...

work in progress...

BRITISH & FOREIGN BIBLE SOCIETY

Jon Riding

Linguistic Computing
at
British & Foreign Bible Society

jon.riding@biblesociety.org.uk
http://lc.bfbs.org.uk